PSEUDO-REPLICATION IN THE NCHS HEALTH EXAMINATION SURVEY

Walt R. Simmons and James T. Baird, Jr.
National Center for Health Statistics

1. *Introduction*

The activities which we shall describe had their origin in the question, "How should one analyze data from a complex survey?" An alternative and perhaps more specific phrasing might be "What are appropriate statistical mechanics for drawing inferences from data secured in stratified multistage probability surveys of social, demographic, or health matters?" It can be assumed that the complex survey design will be accompanied by an elaborate estimation scheme, developed within the context of finite sampling theory. The complex design and elaborate estimation generate inference problems that are different from those solved by most classical statistical analytic techniques.

Rather surprisingly, until a few years ago, this very significant issue received little attention. More recently, its importance is being recognized. See for example, Gurney (1962); McCarthy, Simmons and Losee (1965); Kish (1968); McCarthy (1968); and other references listed in the bibliography. With reluctance, we forego in the present paper any further general discussion of this fascinating question, and restrict ourselves to reporting on some practices and investigations in NCHS that are relevant to the inference problem in complex surveys. We do invite you to keep in mind, however, as we proceed, Kendall's (1961) admonition, "It will be evident that if a sample is not random and nothing precise is known about the nature of the bias operating when it was chosen, very little can be inferred from it about the parent population." And we call attention to the implication that if the variance of a sample statistic is unknown or poorly estimated, the corresponding parameter estimate may have but trivial, and at best ambiguous, value.

2. *Balanced Half-Sample Pseudo-Replication*

Before turning to an account of the way in which replication is being used in the NCHS, a very condensed synopsis is offered of theoretical work in this area carried out in the last few years by Professor Philip J. McCarthy of Cornell University, under a contractual arrangement with our Center. This work is described in greater detail in McCarthy (1966), McCarthy (1968), and in another document not yet in print.

The balanced half-sample pseudo-replication estimator of variance is described at length in an NCHS (1966) publication written by McCarthy. The essential features of the basic half-sample estimator are these:

$\bar{x}'$ is the parent sample estimator of population parameter x

$\bar{y}'_\alpha$ is an estimator of $\bar{x}$ utilizing data from only 1 of the 2 PSU's in each stratum.

$\bar{x}' = \frac{1}{2}(\bar{y}'_\alpha + \bar{y}'_\beta)$, where $\bar{y}'_\beta$ is the complement estimate to $\bar{y}'_\alpha$; i.e., is formed from all PSU's which are in the parent sample but not in the α-half-sample.

$\bar{y}'_\alpha$ and $\bar{y}'_\beta$ are statistically independent

$$E\bar{y}'_\alpha = E\bar{x}' = \bar{x}.$$

The estimator of variance is

$$s^2_{\bar{x}'} = \frac{1}{\lambda}\sum_{\alpha=1}^{\lambda}(\bar{y}'_\alpha - \bar{x}')^2$$

where λ is the number of half-sample replicates utilized.

McCarthy's research has led to a number of findings, and in particular to these conclusions.

A. For estimating the variance of a statistic which is a linear function of sample observations, it is feasible to form pseudo-replicates in such a fashion that a set of a modest number of replications produces not only an unbiased estimate of the true variance, but is identically equal to the value which would be secured if all possible pseudo-replications were formed. (For example, in a 27-strata, 2 PSU's per stratum design, a controlled set of 28 pseudo-replications yields the precise numerical result that could be obtained from the 2^{27} possible replications.)

B. The half-sample replication variance process for only the linear case is shown rigorously to be unbiased. But, very significantly, it is biased only unimportantly for a large class of statistics R (including, for example, a ratio or regression coefficient) if the expected value $E(\bar{r}-\hat{R})^2$ is satisfactorily small, where $\hat{R}$ is the parent sample estimator of R, and $\bar{r}$ is the mean value of the half-sample replicate estimators of R.

C. Estimates of variances developed by the half-sample replication method may be used in the construction of modified tests of hypotheses fully appropriate to the complex design. For example, a pseudo-Chi-square statistic can be calculated to test for independence in a two-way table of estimates, and a modified sign test, using the replicate and complement replicates, provides a low-power, but widely applicable instrument for analysis.

3. *The Health Examination Survey (HES)*

The HES is a major activity of the NCHS in which, through direct examination of a probability sample of the civilian noninstitutional population of the United States, the distributions relating to that population are secured for a considerable variety of physical and physiological characteristics, and prevalences determined for selected medical and dental conditions. The survey design incorporates such features as deep stratification, multistage clustering, controlled selection, adjustments for nonresponse, ratio estimation, and poststratification. One cycle of the survey covers a specified age range of the population and consists of examination of about seven thousand persons. The HES has been described in greater detail in several other reports, and particularly in references NCHS (1965) and NCHS (1967).

Since the cost per examined person is high in the survey, considerable effort is devoted to extracting a maximum amount of information from the data. This implies a requirement for appropriate determination of the sampling variability of a very large number of derived statistics. Computation of variance on an *ad hoc* basis by conventional techniques is practically impossible because of the extreme complexities of relevant algebraic expressions, and the large volume of computations that would be required.

4. *Estimating Variances of Aggregates and Ratios*

We describe first a standardized program in routine use in HES, whereby, using the half-sample pseudo-replication technique, variances are secured for the characteristics under study as an ancillary output to the estimates themselves. For each cell of a table, this procedure typically provides an estimate of a denominator statistic such as number of persons of a specified description, a numerator statistic such as the aggregate value of a measurement for the specified denominator class, and the ratio of numerator to denominator; it produces sampling errors for the denominator, numerator, and ratio; and it can yield for the analyst's convenience supplementary data on sample size, row percentage distributions, or age adjusted estimates.

Specifications for a particular run can be set up by an analyst in one to ten minutes. After the data have been laundered and assembled on basic tapes, running time with current programming and equipment is about 15 minutes per basic table.

An illustration of the process is given in the series of tables A-1 through A-12. These tables show the numbers of decayed, missing, and filled (DMF) teeth of adults classified by age.

Table A-1 gives the numerator under study for each cell in the table, inflated (weighted) to the estimated totals for dentulous males in the United States. For example, it shows that dentulous males age 18-24 in the United States have an estimated 15,322,000 decayed teeth—a resounding and no doubt significant statistic to dental researchers.

Table A-2 gives the corresponding weighted estimates for the denominator for each cell in the table. It shows an estimated 7,022,000 males aged 18-24. In the example this figure is the same for each column of the spread since all of the averages or ratios to be calculated have the same number of persons in the age-sex group as a common denominator. However, if the spread had been specified as some person characteristic (for example economic status), and the statistic being calculated were proportion of persons in each economic group with heart disease, say, the frequencies would, of course, vary for the spread across any given row.

Table A-1. (Numerator) Estimated DMF teeth

Sex and age	DMF	Decayed	Missing	Filled
Grand total--	1,619,629	129,096	853,831	636,719
Total, male-----	755,812	65,242	403,460	287,127
18-24 years-	93,854	15,322	32,339	46,192
25-34 years-	157,852	17,568	62,414	77,869
35-44 years-	184,498	14,042	86,714	83,741
45-54 years-	143,848	10,329	87,478	46,058
55-64 years-	99,529	5,364	71,637	22,527
65-74 years-	60,811	2,152	49,166	9,492
75-79 years-	15,416	462	13,708	1,245
Total, female---	863,816	63,854	450,370	349,591
18-24 years-	117,418	16,714	41,268	59,435
25-34 years-	185,628	17,481	78,942	89,204
35-44 years-	207,281	14,177	101,312	91,791
45-54 years-	164,654	9,609	95,959	59,085
55-64 years-	110,253	3,948	74,392	31,913
65-74 years-	66,326	1,594	48,660	16,071
75-79 years-	12,254	329	9,834	2,090

Table A-3. Average teeth per person

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	17.85	1.42	9.41	7.01
Total, male---	17.19	1.48	9.17	6.53
18-24 years-----	13.36	2.18	4.60	6.57
25-34 years-----	15.77	1.75	6.23	7.78
35-44 years-----	17.23	1.31	8.10	7.82
45-54 years-----	17.99	1.29	10.94	5.76
55-64 years-----	20.44	1.10	14.71	4.62
65-74 years-----	22.33	.79	18.05	3.48
75-79 years-----	24.37	.73	21.67	1.96
Total, female-	18.47	1.36	9.63	7.47
18-24 years-----	14.14	2.01	4.97	7.16
25-34 years-----	17.51	1.64	7.44	8.41
35-44 years-----	18.75	1.28	9.16	8.30
45-54 years-----	19.64	1.14	11.44	7.04
55-64 years-----	21.88	.78	14.76	6.33
65-74 years-----	22.84	.54	16.76	5.53
75-79 years-----	25.02	.67	20.08	4.26

Table A-2. (Denominator) Estimated persons

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	90,713	90,713	90,713	90,713
Total, male---	43,951	43,951	43,951	43,951
18-24 years-----	7,022	7,022	7,022	7,022
25-34 years-----	10,007	10,007	10,007	10,007
35-44 years-----	10,705	10,705	10,705	10,705
45-54 years-----	7,992	7,992	7,992	7,992
55-64 years-----	4,869	4,869	4,869	4,869
65-74 years-----	2,722	2,722	2,722	2,722
75-79 years-----	632	632	632	632
Total, female-	46,761	46,761	46,761	46,761
18-24 years-----	8,300	8,300	8,300	8,300
25-34 years-----	10,597	10,597	10,597	10,597
35-44 years-----	11,050	11,050	11,050	11,050
45-54 years-----	8,383	8,383	8,383	8,383
55-64 years-----	5,037	5,037	5,037	5,037
65-74 years-----	2,902	2,902	2,902	2,902
75-79 years-----	489	489	489	489

Table A-4. Standard deviation of numerator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	21,951	5,045	19,009	16,204
Total, male---	11,589	2,728	10,030	6,914
18-24 years-----	2,403	1,137	1,511	1,945
25-34 years-----	3,982	741	2,713	3,160
35-44 years-----	2,934	1,081	3,558	3,326
45-54 years-----	3,069	1,170	3,711	3,500
55-64 years-----	4,972	737	3,746	2,357
65-74 years-----	3,599	247	3,264	1,925
75-79 years-----	3,184	127	3,065	526
Total, female-	16,745	3,344	12,503	11,187
18-24 years-----	3,143	1,077	1,476	2,826
25-34 years-----	5,485	1,265	3,382	4,084
35-44 years-----	5,041	1,354	2,935	3,179
45-54 years-----	4,592	470	4,128	1,606
55-64 years-----	5,543	505	4,463	2,692
65-74 years-----	4,955	236	3,786	2,151
75-79 years-----	2,635	95	2,254	811

Table A-5. Standard deviation of denominator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	565	565	565	565
Total, male---	372	372	372	372
18-24 years-----	50	50	50	50
25-34 years-----	54	54	54	54
35-44 years-----	81	81	81	81
45-54 years-----	171	171	171	171
55-64 years-----	201	201	201	201
65-74 years-----	141	141	141	141
75-79 years-----	127	127	127	127
Total, female-	496	496	496	496
18-24 years-----	42	42	42	42
25-34 years-----	91	91	91	91
35-44 years-----	126	126	126	126
45-54 years-----	188	188	188	188
55-64 years-----	257	257	257	257
65-74 years-----	191	191	191	191
75-79 years-----	92	92	92	92

Table A-3 presents the calculated ratios or means which are the frequencies in Table A-1 divided by the corresponding frequencies in Table A-2. In this example the data in Table A-3 probably form the heart of the analysis from a subject matter standpoint. In other cases as, for example, the estimated proportion of persons with arthritis, the estimated total number of such persons (Table A-1—weighted numerator) would also be of interest.

Table A-4 presents the sampling errors of the estimates developed in Table A-1, i.e., the standard errors of the national estimates for the numerators.

Table A-5 contains corresponding estimates of the sampling errors of the denominator estimates shown in Table A-2.

Table A-6 presents the estimated standard error for each of the means estimated in Table A-3. This is probably the table that would be most used in evaluating observed trends in this analytical situation. Large sample normal theory would be applied for appropriate testing of hypotheses and calculation of confidence intervals.

Tables A-7 through A-12 are self-explanatory and represent auxiliary output which can be helpful to the analyst.

Table A-7. Age-sex adjusted means or ratios

Sex and age	DMF	Decayed	Missing	Filled
Total-----	17.85	1.42	9.41	7.01
Male-----------	17.19	1.48	9.17	6.53
Female----------	18.47	1.36	9.63	7.47

Table A-6. Standard deviation of ratio

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.22	.05	.20	.17
Total, male---	.21	.06	.21	.14
18-24 years-----	.33	.16	.21	.26
25-34 years-----	.38	.07	.24	.33
35-44 years-----	.24	.09	.34	.28
45-54 years-----	.36	.13	.29	.51
55-64 years-----	.46	.13	.57	.37
65-74 years-----	.45	.06	.95	.66
75-79 years-----	1.28	.19	1.60	.93
Total, female-	.27	.07	.23	.22
18-24 years-----	.35	.13	.17	.32
25-34 years-----	.52	.12	.34	.36
35-44 years-----	.37	.11	.21	.29
45-54 years-----	.36	.05	.40	.18
55-64 years-----	.42	.11	.56	.42
65-74 years-----	.46	.09	.67	.53
75-79 years-----	1.50	.17	1.99	1.38

Table A-8. Numerator sample frequencies (persons)

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	5,483	5,483	5,483	5,483
Total, male---	2,587	2,587	2,587	2,587
18-24 years-----	403	403	403	403
25-34 years-----	662	662	662	662
35-44 years-----	656	656	656	656
45-54 years-----	431	431	431	431
55-64 years-----	262	262	262	262
65-74 years-----	139	139	139	139
75-79 years-----	34	34	34	34
Total, female-	2,896	2,896	2,896	2,896
18-24 years-----	524	524	524	524
25-34 years-----	697	697	697	697
35-44 years-----	702	702	702	702
45-54 years-----	552	552	552	552
55-64 years-----	267	267	267	267
65-74 years-----	132	132	132	132
75-79 years-----	22	22	22	22

Table A-9. Denominator sample frequencies

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	5,483	5,483	5,483	5,483
Total, male---	2,587	2,587	2,587	2,587
18-24 years-----	403	403	403	403
25-34 years-----	662	662	662	662
35-44 years-----	656	656	656	656
45-54 years-----	431	431	431	431
55-64 years-----	262	262	262	262
65-74 years-----	139	139	139	139
75-79 years-----	34	34	34	34
Total, female-	2,896	2,896	2,896	2,896
18-24 years-----	524	524	524	524
25-34 years-----	697	697	697	697
35-44 years-----	702	702	702	702
45-54 years-----	552	552	552	552
55-64 years-----	267	267	267	267
65-74 years-----	132	132	132	132
75-79 years-----	22	22	22	22

Table A-10. Rel-variance of numerator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.00018	.00153	.00050	.00065
Total, male---	.00024	.00175	.00062	.00058
18-24 years-----	.00066	.00551	.00218	.00177
25-34 years-----	.00064	.00178	.00189	.00165
35-44 years-----	.00025	.00593	.00168	.00158
45-54 years-----	.00046	.01284	.00180	.00577
55-64 years-----	.00250	.01888	.00273	.01095
65-74 years-----	.00350	.01324	.00441	.04116
75-79 years-----	.04266	.07564	.04999	.17852
Total, female-	.00038	.00274	.00077	.00102
18-24 years-----	.00072	.00416	.00128	.00226
25-34 years-----	.00087	.00524	.00184	.00210
35-44 years-----	.00059	.00912	.00084	.00120
45-54 years-----	.00078	.00240	.00185	.00074
55-64 years-----	.00253	.01637	.00360	.00712
65-74 years-----	.00558	.02202	.00605	.01792
75-79 years-----	.04625	.08466	.05254	.15069

Table A-11. Rel-variance of denominator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.00004	.00004	.00004	.00004
Total, male---	.00007	.00007	.00007	.00007
18-24 years-----	.00005	.00005	.00005	.00005
25-34 years-----	.00003	.00003	.00003	.00003
35-44 years-----	.00006	.00006	.00006	.00006
45-54 years-----	.00046	.00046	.00046	.00046
55-64 years-----	.00171	.00171	.00171	.00171
65-74 years-----	.00271	.00271	.00271	.00271
75-79 years-----	.04093	.04093	.04093	.04093
Total, female-	.00011	.00011	.00011	.00011
18-24 years-----	.00003	.00003	.00003	.00003
25-34 years-----	.00007	.00007	.00007	.00007
35-44 years-----	.00013	.00013	.00013	.00013
45-54 years-----	.00050	.00050	.00050	.00050
55-64 years-----	.00262	.00262	.00262	.00262
65-74 years-----	.00436	.00436	.00436	.00436
75-79 years-----	.03539	.03539	.03539	.03539

Table A-12. Rel-variance of ratio

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.00016	.00149	.00048	.00060
Total, male---	.00016	.00164	.00054	.00049
18-24 years-----	.00062	.00560	.00223	.00166
25-34 years-----	.00061	.00167	.00155	.00187
35-44 years-----	.00020	.00577	.00186	.00133
45-54 years-----	.00040	.01118	.00071	.00812
55-64 years-----	.00052	.01595	.00154	.00647
65-74 years-----	.00042	.00673	.00280	.03634
75-79 years-----	.00278	.06812	.00548	.22774
Total, female-	.00022	.00297	.00059	.00088
18-24 years-----	.00064	.00441	.00122	.00209
25-34 years-----	.00092	.00580	.00210	.00193
35-44 years-----	.00041	.00836	.00056	.00125
45-54 years-----	.00035	.00208	.00122	.00067
55-64 years-----	.00037	.02137	.00144	.00456
65-74 years-----	.00041	.02915	.00162	.00946
75-79 years-----	.00363	.06928	.00985	.10627

The tabulation described has, of course, undergone constant improvement as experience has been gained and will undoubtedly continue to do so in the future. Since the procedure was first adopted as a "production-type" program in mid-1964, several thousand such tabs have been produced. That the procedure has come to be considered a routine, convenient, and readily available analytical aid, is a positive and encouraging commentary on the applicability of computer support to statistical analysis, as well as the appropriateness and feasibility of the half-sample replication technique of variance estimation.

5. *Further Considerations in Estimating Variances*

The standard procedures just described handle very well the principal HES requirements for appropriate variances of aggregates, means, and ratios. There are many collateral problems, only a few of which can be treated here.

One group of questions relates to cost. Although we were pleased to point out that running time is only about 15 minutes for a set of tables such as A-1 through A-12, computation is still quite expensive for a large number of problems. It is believed that more powerful computers just now becoming available to the Center, along with some reprogramming, will cut costs. But it is reasonable to ask if there may not be less expensive forms of computation which are sufficiently close approximations to the full pseudo-replication technique.

Another class of problems is the extension of the pseudo-replication procedure to estimating variances of statistics other than aggregates, means, or ratios—for example, to the statistics of multiple-regression analysis.

6. *Multiple-Regression Analysis of Anthropometric Data*

In order to study both of these problems—less expensive approximating computations, and the extensions of the replication technique to other statistics—the Survey Research Center at the University of Michigan, in accordance with NCHS specifications, in a project directed by Leslie Kish, prepared for the Center an elaborate series of tables which provide methodological information about multiple regression among anthropometric measurements from the HES. These tabulations derived 16 multiple regression equations, with sampling variance being calculated for regression and correlation coefficients for each by simple random classical methods and by three versions of pseudo-replication, all carried out with three different weighting schemes (approximating devices). Again we must condense the descriptions of our investigations and offer a selection from among the findings.

Each equation studied treated one of 16 body measurements as the dependent variable in a linear regression equation with height, weight, and age as the independent, or predicting, variables. Solutions to these 16 regression problems give the statistics shown in Table B-1 as estimates of the corresponding parameters for the United States, male, noninstitutional population 18-79 years of age. The variables are identified in Table B-2.

We dispose quickly of the matter of three different versions of pseudo-replication—which are described in McCarthy (1966). These versions are labeled, "Basic," "Complement," and "Difference." Variances were calculated by all three methods, but differences among the three for a given weighting scheme are so small, that any one of the three is equally acceptable as an estimator. Indeed, if the estimators had been linear the three should have given identical results. The closeness of the three values is a comforting piece of evidence that the non-linearity in the present case is not importantly troublesome. Both this result and the validity of the pseudo-replication technique for non-linear statistics such as regression and correlation coefficients are assured if the mean of the replicate statistics is closely equal to the corresponding statistic in the parent sample. For the anthropometric data this condition was fully satisfied. For all variables and all correlation coefficients the largest mean discrepancy was that between the body measurements and age, and this was only 0.0016 with a standard deviation of 0.0014 among the 16 body measurements. Contrastingly, Table B-3 shows substantial differences between "assumed simple random" and the half-sample methods, for all weighting schemes, and this result is typical for the full set of body measurements.

Weighting Scheme I disregards the actual complex survey design and estimating procedure, and treats the data as though they were a simple random sample—i.e., each case is given unit weight.

In the Health Examination Survey three sets of adjustment factors are applied in addition to the reciprocal of the probabilities of selection, in order to take advantage of ratio estimation, poststratification, and imputation for nonresponse. Some of these calculations are made specific for defined tabulation areas in the United States, and while straightforward, involve considerable work. In the full pseudo-replication procedure these adjustment factors are calculated and applied specific for each half-sample. The process is identified here as Weighting Scheme III, or as "unique weight." When the adjustment factors of

Table B-1. Test statistics from multiple linear regression of each of 16 body measurements on age, height, and weight

Statistics	Number of tests
Mean	19
Simple correlation coefficient	51
Partial correlation coefficient	48
Multiple correlation coefficient	16
Beta coefficient--age	16
Beta coefficient--height	16
Beta coefficient--weight	16

Table B-2. Variables used in regression analyses

Independent	Dependent
Age	Biacromial diameter
Height	Right arm girth
Weight	Chest girth
	Waist girth
	Right arm skinfold
	Infrascapular skinfold
	Sitting height--normal
	Sitting height--erect
	Knee height
	Popliteal height
	Thigh clearance height
	Buttock-knee length
	Buttock-popliteal length
	Seat breadth
	Elbow-elbow breadth
	Elbow rest height

Table B-3. Illustrative comparison of variances for alternative weighting schemes and methods of estimating variance

[Mean male chest girth measured in cm.]

Methods of estimating variance and rel-variance	Weighting Scheme		
	I Unit weight	II Con-stant weight	III Unique weight
	(Rel-variance times 10^5)		
Assumed simple random sampling	.2272	.2284	.2284
Basic half-sample	.3625	.3771	.3256
Complement half-sample	.3610	.3977	not cal-culated
Difference half-sample	.3616	.3866	not cal-culated

Table B-4. Percent distribution of sample persons by relative final weight

Relative weight	Percent distri-bution
All weights	100.0
1/2	13.8
1	58.6
2	16.4
3	9.3
4-7	1.9

the parent or whole sample are applied unchanged to each half-sample, the process is called Weighting Scheme II, or "constant weight." This process was tested to obtain an indication as to whether the bias introduced by such a procedure is of practical significance. Weighting Scheme II involves much less expense than Weighting Scheme III, and would be preferred if it did not introduce significant bias. For the test data presented in this paper, Weighting Scheme II encompasses a fully balanced set of half-sample replications. Weighting Scheme III has incomplete balancing as a result of calculations which were not fully faithful to the design.

7. *Impact of Weighting on Estimates of Primary Statistics*

The tabulations provide empirical evidence on several aspects of estimation or weighting effects, and in particular on two classes of problems. One class consists of comparisons between the unweighted (biased) versions and the properly weighted values for primary statistics from the parent sample such as the mean, and correlation and regression coefficients. The second class consists of comparisons between variances for these same statistics computed on the one hand under the assumption of simple random design, and on the other by pseudo-replication using either Weighting Scheme II or III.

The extent of deviation from equal-weighting brought about by all steps of sampling and estimating is summarized in Table B-4 which gives the percentage distribution of sample persons by approximate size of final relative weight.

Table C-1 presents for the anthropometric data the distributions of ratios of unweighted to weighted values for the primary statistics.

The data clearly indicate, even for this design in which there is some approximation to equal weighting as indicated by the distribution in Table B-4, that numerous cases exist for which one does ***not*** obtain reasonably good approximations to the estimates by the use of "unweighted" data. Means were relatively stable, although a difference of one full year in mean age, for example—this was one of the observations—while not large percentagewise, is of considerable practical significance. Discrepancies larger than 10 percent were obtained for 18 percent of the linear correlation coefficients calculated. The coefficients of multiple correlation showed a possibly unexpected stability, with all "unweighted" coefficients having errors of five percent or less. Other statistics manifested expected patterns of substantial variation

Table C-1. Distributions of ratios of "unweighted" to weighted estimates, selected primary statistics, HES anthropometric data

Ratio	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
0.0-0.24							
0.25-0.49			1				
0.50-0.74		3	1		1		
0.75-0.89		2	4		1	1	
0.90-0.94		4	2		3	1	
0.95		3	3		1		
0.96		1	1	1		2	2
0.97	1	3	5			3	
0.98	2	6	3	3		5	3
0.99	1	4	4	4		1	1
1.00	15	10	8	7	2	1	2
1.01		5					4
1.02			1				
1.03		2	1			1	1
1.04		1	1		2		
1.05		1	2	1			1
1.06-1.10		2	4		1		
1.11-1.25		2	2		1	1	2
1.26-1.50		2	1		1		
1.51-1.75							
1.76-2.00							
Over 2.00			4		3		
Number of measures	19	51	48	16	16	16	16
Median ratio	0.999	0.994	0.995	0.995	1.00	0.977	1.00
Mean ratio	0.996	0.985	1.99	0.994	3.91	0.979	1.08
Standard deviation	0.009	0.133	5.69	0.018	9.47	0.068	0.225

between weighted and "unweighted" estimates which would tend to discourage use of the latter to approximate the former, except perhaps for an occasional preliminary or rough approximation. The causes of these differences have their source in a variety of circumstances. While we can isolate some of the factors—e.g., the difference in mean age is partly the consequence of differential nonresponse by age—the real moral to be drawn is "weight properly," and avoid the risk of unpredictable bias which may be introduced if weighting is suppressed.

8. *Variances of Primary Statistics*

The relationship of variances appropriate to the design and estimation procedure calculated by the replicate method, to those obtained by use of a simple random sampling formula have several important aspects. We shall concentrate on two of these but will note others briefly. The strategy of our approach to analysis of this matter calls for comparison of the fully balanced replication procedure with the "unweighted" simple random variances in Weighting Scheme I. We did not have the precise data preferred for this comparison. Weighting Scheme III had the full weighting methodology, but was incompletely balanced since it used a collapsed grouping of 16 strata rather than the more nearly ideal combination which treated random halves of the certainty strata as PSU's. The variances in Weighting Scheme II carried complete balancing, but used the "constant" weights as defined above, rather than the conceptually superior "unique" weights. Given this dilemma, we chose a computational course which compared first the variances under Weighting Scheme II with the simple variances of Weighting Scheme I, followed by comparison of Weighting Schemes II and III with one another.

If the ratio of replicate variance to simple random variance is near unity, the latter may serve adequately since it is much less costly, and its own sampling error is smaller. Even if the ratio were not near unity, but were approximately constant, that would permit use of the simple random sampling formula multiplied by a constant as a probably satisfactory measure of precision.

The data in Table D-1 indicate, however, not only that the simple random variances are generally too small, but that the range of ratios of proper to simple random variances show such wide ranges of variability that application of a constant "correction factor" to the unweighted simple random sample estimate of variance would not be a practical procedure for the test data studied. While there is some tendency for the ratios of correlation statistics to center about values in the range 1.5-1.8, and at 3.2 for the mean variances, the variability from one body measurement to another, is clearly too great.

There is the possibility that the unsatisfactory result displayed in Table D-1 is the consequence of *different weighting* of observations in the two estimating methods rather than in the variance formulae which distinguish between the actual design and the assumed simple random design. To explore this possibility, ratios of proper variances to simple variances were calculated, with individual observations in both numerator and denominator carrying the same weight. The distributions of ratios are shown in Table D-2. Patterns for all statistics remain essentially unchanged from those observed in Table D-1, and result in a corresponding conclusion—for the empirical situation under study a constant "correction factor" applied to the weighted simple random sampling variance estimates is not acceptable.

Table D-3 compares two sets of replicate variances, one using the constant weights of the parent sample and the other, unique weights for each replicate. The impact of this difference in weighting, when both methods employ a proper variance formula, is more difficult to evaluate. The means and medians of ratios of the two variances are near unity, and the average of the mean ratios is almost precisely unity. The variances of these ratios among the statistics are smaller than those observed in the other variance comparisons—thus indicating that the simple Weighting Scheme II might be acceptable. But the variability of the ratios still is not trivial, and on this point further calculations for a wider range of statistics are in order. Data from Tables D-1 through D-3 are recapitulated in Table D-4 in a manner which focuses attention on highlights. In condensed summary, those highlights are

1. The ordinary simple variance formula yields unsatisfactory approximations to true variance in this complex survey, whether original data are used in weighted or unweighted form. (1st two data columns, Table D-4)

Table D-1. Distributions of ratios of Weighting Scheme II replicate half-sample variance to simple random sample Weighting Scheme I variance, HES anthropometric data

Ratio	Primary statistics to which variances relate						
	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
0.51-1.00		5	2	2	2		1
1.01-1.50	1	19	17	4	6	6	5
1.51-2.00	2	15	14	4	6	7	7
2.01-2.50	3	7	7	1	2	2	1
2.51-3.00	2	4	4	1		1	2
3.01-3.50	4		2	2			
3.51-4.00	3						
4.01-4.50	1		1	1			
4.51-5.00	1	1	1				
5.01-5.50				1			
5.51-6.00	2						
Number of measures	19	51	48	16	16	16	16
Median ratio	3.19	1.55	1.68	1.75	1.51	1.65	1.65
Mean ratio	3.26	1.68	1.87	2.16	1.50	1.69	1.69
Standard deviation	1.22	0.689	0.805	1.24	0.433	0.428	0.527

Table D-2. Distributions of ratios of replicate variances to simple variances, all data weighted by Scheme II, selected statistics, HES anthropometric data

Ratio	Primary statistics to which variances relate						
	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
Under 0.51							
0.51-1.00		4	2	2	2		1
1.01-1.50	1	18	17	4	4	5	7
1.51-2.00	2	16	10	3	7	9	5
2.01-2.50	3	8	12	2	3	1	2
2.51-3.00	2	3	2	1		1	1
3.01-3.50	4	1	3	1			
3.51-4.00	3		1	2			
4.01-4.50	1						
4.51-5.00	1	1	1				
5.01-5.50	1			1			
5.51-6.00	1						
Over 6.00							
Number of measures	19	51	48	16	16	16	16
Median ratio	3.19	1.62	1.76	1.84	1.65	1.67	1.50
Mean ratio	3.23	1.73	1.90	2.19	1.60	1.69	1.60
Standard deviation	1.18	0.696	0.790	1.22	0.458	0.390	0.491

Table D-3. Distributions of ratios of replicate variance Weighting Scheme II to replicate variance Weighting Scheme III, selected statistics, HES anthropometric data

Ratio	Mean	Correlation coefficient: Simple	Correlation coefficient: Partial	Correlation coefficient: Multiple	Regression coefficient on: Age	Regression coefficient on: Height	Regression coefficient on: Weight
0.0-0.24			2		1		1
0.25-0.49	1				1		
0.50-0.74		7	7	5	1	2	1
0.75-0.89	1	10	15	4	4	4	4
0.90-0.94	3		4			1	4
0.95-1.05	4	7	6	1	1	4	1
1.06-1.10	3	5	3	1	2	1	3
1.11-1.25	4	7	7	2	3	2	2
1.26-1.50	1	8	2	1	2	1	
1.51-1.75	1	4	1	1	1	1	
1.76-2.00	1						
Over 2.00		3	1	1			
Number of measures	19	51	48	16	16	16	16
Median ratio	1.06	1.07	0.895	0.857	0.975	0.975	0.920
Mean ratio	1.09	1.15	0.932	1.02	0.966	0.995	0.896
Standard deviation	0.304	0.473	0.327	0.472	0.373	0.255	0.248

Table D-4. Recap of highlights from Tables D-1 through D-3

Type of statistic	Median ratios of estimates of variance: Correct to unweighted simple random variances	Median ratios of estimates of variance: Correct to weighted simple variances	Standard deviations of ratios of variances over all test variables: Correct to unweighted simple random variances	Standard deviations of ratios of variances over all test variables: Correct to weighted simple variances	Standard deviations of ratios of variances over all test variables: Replicate Weighting Scheme II to replicate Weighting Scheme III[1]
Mean	3.19	3.19	1.22	1.18	0.30
Correlation coefficient					
Simple	1.55	1.62	0.69	0.70	0.47
Partial	1.68	1.76	0.80	0.79	0.33
Multiple	1.75	1.84	1.24	1.22	0.47
Regression coefficient					
Age	1.51	1.65	0.43	0.46	0.37
Height	1.65	1.67	0.43	0.39	0.26
Weight	1.65	1.50	0.53	0.49	0.25

[1]The overall average ratio for this comparison is 1.007.

2. Even multiplication of simple variances by a calibrating constant would not produce adequate measures of sampling variance. (Note the substantial variations shown in 3rd and 4th data columns of Table D-4.)

3. In both the above comparisons, whether the data are weighted or unweighted makes little difference in the estimated variance.

4. When the pseudo-replication method is employed, it is unclear whether the "constant" weights of Scheme II are satisfactory approximations for the "unique" weights of Scheme III. The simpler scheme appears to be unbiased, but may understate or overstate variance too frequently. (Data column 5 and the footnote of Table D-4)

9. Addenda

9.1 *Addendum I.* NCHS at present has in routine use only the programs described earlier for handling aggregates and means or ratios, but there should be available in the fairly near future a pseudo-replication multiple regression analysis package and several other statistical measures which are being programmed by SRC, Michigan, under contract with the Center.

9.2 *Addendum II.* It may be that the most significant feature of the pseudo-replication process has not yet been exploited. Consider once more the 2-way table of DMF teeth. As a step in the process of calculating variances, the program builds quite literally 28 half-sample replicates of that parent table. Each of these half-sample replicates is an estimator of the same statistics as appear in the parent table. The variance among the replicates *is* the variance of their mean value. At the analyst's option each of the replicates can be printed out. When that is done, there is available not just a mean and a variance, but a *distribution* of 28 correlated estimates. Indeed the analyst now has before him 28 pseudo-replications of an experiment, as it were. Thus far, we have used this wealth of material only occasionally, and in unstructured ways. (The print-outs have been used also as the vehicle for calculating variance of medians and other position statistics.) But we suspect that the replicated half-sample tables contain building blocks that can become the basis for significant analytical structures of a new order. We commend study of this intriguing possibility to researchers.

REFERENCES

Gurney, Margaret, "The Variance of the Replication Method for Estimating Variances" (Unpublished memorandum, U.S. Bureau of the Census, Washington, D.C., 1962).

Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics,* Vol. II, p. 1, Hafner Publishing Co., New York, 1961.

Kish, Leslie, "Design and Estimation for Subclasses, Comparisons, and Analytical Statistics," paper presented at the Symposium on Foundations of Survey Sampling, Chapel Hill, N.C., April 22-26, 1968.

McCarthy, Philip J., Simmons, Walt R., and Losee, Garrie J., "Replication Techniques for Estimating Variances from Complex Surveys," paper presented at Joint Session of the Epidemiology and Statistics Sections, APHA, Chicago, Ill., October 18, 1965.

McCarthy, Philip J., *Replication: An Approach to the Analysis of Data from Complex Surveys,* PHS Publication No. 1000, Series 2, No. 14, Washington, D.C., April 1966.

McCarthy, Philip J., "Pseudo-Replication: Half-Samples," paper presented at the Symposium on Foundations of Survey Sampling, Chapel Hill, N.C., April 22-26, 1968.

National Center for Health Statistics, *Plan and Initial Program of the Health Examination Survey,* PHS Publication No. 1000, Series 1, No. 4, Washington, D.C., July 1965.

National Center for Health Statistics, *Plan, Operation, and Response Results of a Program of Children's Examinations,* PHS Publication No. 1000, Series 1, No. 5, Washington, D.C., October 1967.